DISENTANGLING OBSERVATION BIASES TO MONITOR SPATIO-TEMPORAL SHIFTS IN SPECIES DISTRIBUTIONS

Diego Marcos, Christophe Botella, Pierre Alliez & Alexis Joly
Inria Université Côte d'Azur
France
{first.last}@inria.fr

Ilan Havinga
Wageningen University
The Netherlands
{first.last}@wur.nl

Cassio F. Dantas & Dino Ienco
Tetis, INRAE
France
{cassio.fraga-dantas, dino.ienco}@inrae.fr

ABSTRACT

The accelerated pace of environmental change due to anthropogenic activities makes it more important than ever to understand current and future ecosystem dynamics at a global scale. Species observations stemming from citizen science platforms are increasingly leveraged to gather information about the geographic distributions of many species. However, their usability is limited by the strong biases inherent to these community-driven efforts. These biases in the sampling effort are often treated as noise that has to be compensated for. In this project, we posit that better modelling the sampling effort (including the usage of the different platforms across countries, local accessibility, attractiveness of the location for platform users, affinity of different user groups for different species, etc.) is the key towards improving Species Distribution Models (SDM) using observations from citizen science platforms, thus opening up the possibility of leveraging them to monitor changes in species distributions and population densities.

1 INTRODUCTION

Climate change is set to have dramatic impacts on biodiversity worldwide (Arneth et al., 2020), already driving shifts in species distributions with serious consequences for peoples' quality of life (Shin et al., 2019). Predicting the effect of climate change on biodiversity is therefore an important and highly-active field of research. Predictions can alert stakeholders and decision makers to potential future risks, motivate actions to sustain biodiversity and support the development of proactive Climate Change Adaption policies to reduce climate change impacts (Bellard et al., 2012).

Predicting the impacts of climate change on biodiversity requires accurate Species Distribution Models (SDMs). These models often rely on observations of presence and absence of a species obtained via systematic field surveys as labels and geo-located variables, such as climate data, soil maps or satellite imagery as inputs. However, survey data are very labor intensive, severely limiting their spatial scope and temporal frequency, which prevents to keep pace with the rapid effects of climate change (Jetz et al., 2019) and of other anthropogenic environmental disturbances. For this reason, big citizen science data has become especially relevant in generating predictions using SDMs due the large spatial scope and frequency at which this data are made available when compared to more systematic surveys (Geldmann et al., 2016).

Nevertheless, citizen science data contain very significant biases due to the manner in which the data are generated. These biases are related to the distribution of human population, the accessibility of species' habitats, their seasonality, visibility and charisma (Chandler et al., 2017), and the usage of the citizen science platforms themselves. This has resulted in significant information gaps in biodiversity-rich but data-poor regions and a lack of species-level information for some of the most threatened species and biomes. These biases therefore present a fundamental challenge to the use of citizen science data in SDMs (Amano et al., 2016). Previous research has shown that explicitly accounting for the sampling effort can result in better SDMs (Botella et al., 2021), and our main hypothesis is that building a rich model of the sampling effort is the most promising direction to improve the state-of-the-art in SDMs. Although some previous work exists in modeling user behavior with respect to the time of the year and taxonomic groups (Di Cecco et al., 2021), this project would be the first, to the best of our knowledge, that aims at jointly learning the user behavior and species distributions in an end-to-end manner, thus leveraging the synergies between both tasks.

We aim to develop a relevant new methodology in this under-explored area of research through the use of a hybrid approach that makes use of both 1) the large amounts of available, although biased, observation data and 2) priors about the behavior of citizen scientists and wildlife. The exponential growth in user base that these platforms have experienced in the last few years means that there is currently a unique opportunity to improve SDMs in low data areas by modeling the interactions between human observers and species in addition to their distributions. We will disentangle these important additional drivers of human observation bias to support the next generation of SDMs. Specifically, we will design Bayesian Belief Networks (BBNs) that simultaneously model observer and species densities, along with observer-species affinity estimates, to obtain an informed approximation of the sampling effort (Figure 1). To do this, we will utilize big open access data on species presence available from citizen science platforms iNaturalist and Pl@ntNet, as well as other large photographic repositories such as Flickr, from which species observations can be inferred by applying existing deep learning models for species identification. The use of BBNs will allow us to harness the ever-growing amount of observation data while leveraging prior knowledge about the behavior of both natural species and human observers (Moreira et al., 2021).

Modelling human observers and their affinities for different species in this way will allow us to understand whether a change in the number of species observations in a region can be explained by a change in observer interest or by an actual change in the species population density. We expect this approach to be particularly useful in areas in which a species has recently ceased to exist or has undergone a rapid population decline, thus allowing for early warnings of receding species distributions or population densities. Such findings will help identify key climate change impacts and carry important policy implications. Within this project we will not only aim at disseminating its findings via publications in high impact open access journals and an open source code repository, but also by integrating the results into the Pl@ntNet platform in order to provide better species identification based on geo-location, and by seeking collaborations to apply the developed models to specific case studies.

2 METHOD

This project requires the use of a method that is able to learn from data and incorporate priors about the behavior of the different elements being modeled. For this, we will explore the use of models that combine point processes Cox & Isham (1980), neural networks, and BBNs. In particular, the BBN will encode the priors about the species and human observer distributions and how they relate to each other, while the neural network will learn how to predict both distributions from the input data and map them to suitable probabilities to be fed to the BBN, and the point process will be used model the uploading of observations to the citizen science platforms, see Figure 1. We will seek to jointly model, at every location, the density of observers and their typology (usage of the citizen science mobile app, species and life stages they are most attracted by, etc.) along with the abundance of the considered species, thus going beyond presence/absence binary prediction. For instance, we can model a set of species observations in space and time, in a spatial domain D during a time interval T, based on a Poisson point process (PPP) of intensity $\lambda : D \times T \to \mathbb{R}^+$:

$$Y \sim PPP(\Lambda),\ \ \forall x \in D, t \in T,\ \ \lambda(x,t) = \ S(x,t)\, H(x,t)\, R(x,t). \tag{1}$$

Extending the model of Fithian et al. (2015), S represents the intensity of a PPP modeling the abundance of the species abundance in space and time. The points of this species process represent the

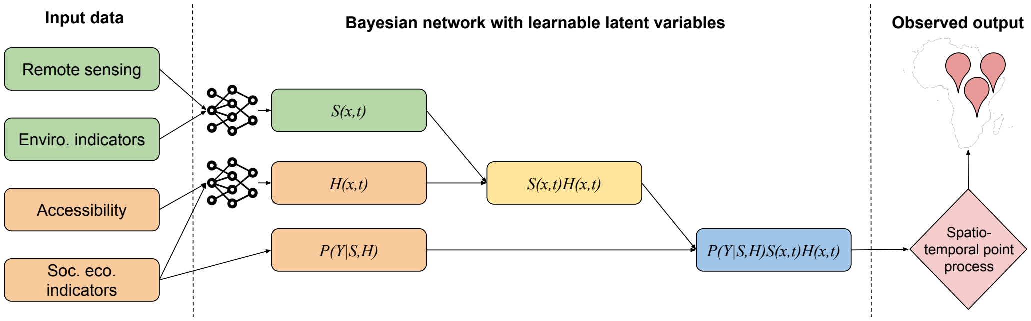

Figure 1: With the aim of learning better species distribution models (SDM), we will set up an end-to-end learnable model that emulates the process that leads to the generation of citizen science observations, which includes the SDM itself but also the distribution of observers and their affinity towards the species. The links that require complex, non-linear functions, such as the prediction of habitat suitability from environmental and remote sensing indicators, will be encoded as neural networks to be learned along with the rest of the model.

species individuals, which are thinned with an observation probability that varies in space and time. This is based on the probability that a human observer is present $H(x, t)$ at a given place and time (sampling effort), and that the species is reported $R(x, t) = P(Y|S(x, t), H(x, t))$ conditionally to the observer presence and species abundance. For instance, H could model the effect of accessibility, population density or seasonality on sampling effort, while R could model the affinity of observers towards that species, and the effect of species abundance and sampling effort on its detectability. The three functions S, H and R could be modelled by a neural network with specific parameters. Note that is just one possible instantiation of the problem and we could also account for dependencies between species. We can then express the PPP negative log-likelihood of the observations $Y = (x_1, t_1), ..., (x_n, t_n)$:

$$-\log(P(Y)) = \int_D \int_T \lambda(x, t)dtdx - \sum_i \log(\lambda(x_i, t_i)),$$

which can be approximated by evaluating the integral only in quadrature points $(x_1^0, t_1^0), ..., (x_Q^0, t_Q^0)$ as in Berman & Turner (1992). In order to help the method solve the ambiguity posed by the symmetry in Eq. (1) between H, R and S we will explore the use of different types of priors, for instance by using input features $\mathbf{z}_h$ and $\mathbf{z}_s$ that are known to be linked to either human and wildlife behaviour respectively, the combination of the presence-only observations with a dataset of systematic surveys of species local presence/absence that has been compiled as part of the GeoLifeCLEF 2023[1] contest, in which some of the authors are involved. A part of this presence-absence dataset will also be used to evaluate our models. Hence, using the improved SDMs provided by the model to be developed in this project, we will move ahead with studying what new possibilities they offer in terms of answering research questions that remain elusive with current SDMs, namely detecting fine-grained spatio-temporal shifts in species distributions and population densities.

3 Ambition

With this research project, we seek to develop methods that will provide insights about (1) the current species distributions and trends, and (2) the way in which citizen scientists interact with these species. This research will help fill the current gap in employing machine learning methods to quantify biases in species observation datasets (Beery et al., 2021). The most relevant studies in the literature have highlighted the existence of different observer-species affinities (Mac Aodha et al., 2019), or explored the influence of contextual information, such as location, at a regional scale (Terry et al., 2020). We will go beyond this by explicitly seeking to quantify observer-species affinities using all the available data from multiple citizen science platforms. Ultimately, our research will strengthen our capacity of generating novel, data-driven discoveries in the field of biodiversity conservation via the valorization of the data generated by biodiversity observation platforms.

[1] To be described in `https://www.imageclef.org/GeoLifeCLEF2023`

REFERENCES

Tatsuya Amano, James DL Lamming, and William J Sutherland. Spatial gaps in global biodiversity information and the role of citizen science. *Bioscience*, 66(5):393–400, 2016.

Almut Arneth, Yunne-Jai Shin, Paul Leadley, Carlo Rondinini, Elena Bukvareva, Melanie Kolb, Guy F Midgley, Thierry Oberdorff, Ignacio Palomo, and Osamu Saito. Post-2020 biodiversity targets need to embrace climate change. *Proceedings of the National Academy of Sciences*, 117 (49):30882–30891, 2020.

Sara Beery, Elijah Cole, Joseph Parker, Pietro Perona, and Kevin Winner. Species distribution modeling for machine learning practitioners: a review. In *ACM SIGCAS Conference on Computing and Sustainable Societies*, pp. 329–348, 2021.

Céline Bellard, Cleo Bertelsmeier, Paul Leadley, Wilfried Thuiller, and Franck Courchamp. Impacts of climate change on the future of biodiversity. *Ecology letters*, 15(4):365–377, 2012.

Mark Berman and T Rolf Turner. Approximating point process likelihoods with glim. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(1):31–38, 1992.

Christophe Botella, Alexis Joly, Pierre Bonnet, François Munoz, and Pascal Monestiez. Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data. *Methods in Ecology and Evolution*, 12(5):933–945, 2021.

Mark Chandler, Linda See, Kyle Copas, Astrid MZ Bonde, Bernat Claramunt López, Finn Danielsen, Jan Kristoffer Legind, Siro Masinde, Abraham J Miller-Rushing, Greg Newman, et al. Contribution of citizen science towards international biodiversity monitoring. *Biological conservation*, 213:280–294, 2017.

David Roxbee Cox and Valerie Isham. *Point processes*, volume 12. CRC Press, 1980.

Grace J Di Cecco, Vijay Barve, Michael W Belitz, Brian J Stucky, Robert P Guralnick, and Allen H Hurlbert. Observing the observers: How participants contribute data to inaturalist and implications for biodiversity science. *BioScience*, 71(11):1179–1188, 2021.

William Fithian, Jane Elith, Trevor Hastie, and David A Keith. Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, 6(4):424–438, 2015.

Jonas Geldmann, Jacob Heilmann-Clausen, Thomas E Holm, Irina Levinsky, BO Markussen, Kent Olsen, Carsten Rahbek, and Anders P Tøttrup. What determines spatial bias in citizen science? exploring four recording schemes with different proficiency requirements. *Diversity and Distributions*, 22(11):1139–1149, 2016.

Walter Jetz, Melodie A McGeoch, Robert Guralnick, Simon Ferrier, Jan Beck, Mark J Costello, Miguel Fernandez, Gary N Geller, Petr Keil, Cory Merow, et al. Essential biodiversity variables for mapping and monitoring species populations. *Nature ecology & evolution*, 3(4):539–551, 2019.

Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9596–9606, 2019.

Catarina Moreira, Yu-Liang Chou, Mythreyi Velmurugan, Chun Ouyang, Renuka Sindhgatta, and Peter Bruza. Linda-bn: An interpretable probabilistic approach for demystifying black-box predictive models. *Decision Support Systems*, 150:113561, 2021.

YJ Shin, A Arneth, RR Chowdhury, F Midgley Guy, E Bukvareva, A Heinimann, AI Horcea-Milcu, M Kolb, P Leadley, T Oberdorff, et al. Chapter 4, plausible futures of nature, its contributions to people and their good quality of life. *IPBES Global Assessment on Biodiversity and Ecosystem Services, Intergovernmental Science Policy Platform on Biodiversity and Ecosystem Services, Bonn, Germany*, 2019.

J Christopher D Terry, Helen E Roy, and Tom A August. Thinking like a naturalist: Enhancing computer vision of citizen science images by harnessing contextual data. *Methods in Ecology and Evolution*, 11(2):303–315, 2020.